**WordCloud HW4**

---

title: "Text mining 1"

author: "Justin Lee"

date: "2018/4/9"

output: html\_document

---

> My first Word Cloud

## 初訪文字探勘

>這是我做text mining word cloud的第一個嘗試。我選定的對象是壹週刊的FBpage，想看壹週刊的臉書貼文是不是符合大家「腥羶色」的印象。於是我利用FB API 爬了300則貼文，並做文字雲。

token<-

'EAACEdEose0cBALwncqM5T5iIpZCdLEjE8WYema9PxSY397Yjc5qHrsDZC4oeOQVJ7gkGJfCZAU6ARW64DZA8TKo6JZCideVrfe7B5UdOENOz6x75ZCUDbLp0EZBuZAkjrs9UB60xODYLZCGZAkoxDjXTBClmB1HbBVXqlZCAIkY7ZApbZB4MeuJNbPo4u9RxfnZAk4bme2sMkuG1tLZBwZDZD'

## token 會變，所以要用的時候需更新

install.packages('Rfacebook')

install.packages('devtools')

library('Rfacebook')

library('devtools')

## Test my token

test<-getUsers('me',token,private\_info=F)

test$name

# Using R facebook

#curl -i -X GET \

#"https://graph.facebook.com/v2.12/nexttw?fields=release\_date%2Cposts&access\_token=EAACEdEose0cBAC8sFeJxZB99wZAZAiRkeMXi6ZBbP6xWr3n9LGYi5mjggWfD17DsJkBbFLVkxpQLxBf12OZBdlz3ZBsgT4yOSpvFCm2InZA5sIUQ3wFstCBZAW17Oo4BZCuZCUGh87jSCAqiNovfqSLlG1wXIUBaQTB14JOe92vhGHBF0220GlX7VPzjX1BqTsEJ62JSvzViZBexgZDZD"

## 壹週刊

page.id<- "125693144152226"

page<-getPage(page.id,token,n=300) #n=300 代表你要取300個posts

View(page[page$likes\_count==max(page$likes\_count),])

write.csv(page,'tm.csv')

```{r install, include=FALSE}

library("tm")

library("SnowballC")

library("wordcloud")

library("RColorBrewer")

library('tmcn')

library('jiebaR')

library('jiebaRD')

```

## 安裝完相關套件後，開始清資料，把對分析不相干的字料清除。例如：stopwords，跟無意義的字串，例如：「嗯」

```{r eval=FALSE, include=FALSE}

page<-read.csv('~/Desktop/tm.csv')

```

```{r run, echo=TRUE}

docs<-Corpus(VectorSource(page$message))

inspect(docs)

tospace<-content\_transformer(function(x,pattern)gsub(pattern,'',x))

docs<-tm\_map(docs,removePunctuation)

docs<-tm\_map(docs,removeNumbers)

docs<-tm\_map(docs,function(word){gsub('[A-Za-z0-9]','',word)})

docs<-tm\_map(docs,tospace,'/')

docs<-tm\_map(docs,tospace,'@')

docs<-tm\_map(docs,tospace,'\\|')

docs<-tm\_map(docs,stripWhitespace)

docs<-tm\_map(docs,tospace,'我')

docs<-tm\_map(docs,tospace,'的')

docs<-tm\_map(docs,tospace,'了')

docs<-tm\_map(docs,tospace,'呵')

docs<-tm\_map(docs,tospace,'呢')

docs<-tm\_map(docs,tospace,'嘛')

docs<-tm\_map(docs,tospace,'是')

docs<-tm\_map(docs,tospace,'啊')

docs<-tm\_map(docs,tospace,'就')

docs<-tm\_map(docs,removeWords,stopwordsCN())

```

> 用jeiba package進行斷詞

```{r jeiba, echo=TRUE}

mixseg = worker()

jieba\_tokenizer=function(d){

unlist(segment(d[[1]],mixseg))

}

seg = lapply(docs, jieba\_tokenizer)

freqFrame = as.data.frame(table(unlist(seg)))

freqFrame<-freqFrame[1:3116,]

View(freqFrame)

pal <- brewer.pal(8, 'Dark2')

wordcloud(freqFrame$Var1,freqFrame$Freq,min.freq = 9,

random.order=F,random.color=T,

rot.per=0.3,colors=pal,

use.r.layout = F,fixed.asp=T,family='Heiti TC Light')

```



從這次分析中可以看出來，壹週刊的FB 貼文中，強調自己的報導很完整，且追戲出現很多遍，而這次我篩選詞彙的標準為Frequency要超過9次以上，因此有稍微符合我們對壹週刊是娛樂性報導的印象，比較沒有政治或經濟的議題。但這次的分析只有擷取300篇貼文，若將擷取提文數量增加到1000篇，可能會有更明顯的發現。

由於我的RMD檔跑不出來不知道為什麼，所以先用Word 檔交作業。